W540–W543 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki478

YASS: enhancing the sensitivity of DNA similarity search

Laurent Noé* **and Gregory Kucherov**

LORIA/INRIA-Lorraine, 615 rue du Jardin Botanique, 54602 Villers-les-Nancy, France

Received March 23, 2005; Accepted April 15, 2005

ABSTRACT

YASS is a DNA local alignment tool based on an efficient and sensitive filtering algorithm. It applies transition-constrained seeds to specify the most probable conserved motifs between homologous sequences, combined with a flexible hit criterion used to identify groups of seeds that are likely to exhibit significant alignments. A web interface (http://www.loria.fr/projects/YASS/) is available to upload input sequences in fasta format, query the program and visualize the results obtained in several forms (dot-plot, tabular output and others). A standalone version is available for download from the web page.

INTRODUCTION

Modern bioinformatics relies heavily on alignment programs and motif discovery tools, and numerous comparative genomics projects need ever more precise and faster tools for comparing two or several genomic sequences with different resolutions.

Except for small sequences, the exact local alignment algorithm of Smith and Waterman (1) is not frequently used, and most alignments are obtained using heuristic alignment tools such as FASTA (2), FLASH (3), BLAST (4,5), BLASTZ (6) and PatternHunter (7,8). All these methods introduce a trade-off between two competing parameters: *selectivity* (or *specificity*) directly affecting the speed of the algorithm and *sensitivity* affecting its precision (i.e. the number of relevant alignments missed). Achieving a good trade-off between sensitivity and selectivity is the key issue in local alignment tools. The recently introduced *spaced seeds* technique (7,8) allows an increase in sensitivity without loss in selectivity. This innovation triggered various studies (9–15) related to the usage, design and generalizations of spaced seeds.

In this note, we present YASS (Yet Another Similarity Searcher)—a new software for computing local alignments of two DNA sequences—and its web server (http://yass.loria.fr/interface.php). Compared with other tools, YASS is based on two innovations. The first is a new spaced seed model called *transition-constrained seeds* that takes advantage of statistical properties of real genomic sequences. The second feature is a new statistically founded *hit criterion* that controls the formation of groups of closely located seeds that are likely to belong to the same alignment. An implementation of these improvements, reported here, provides a fast and sensitive tool for local alignment of large genomic sequences.

DESCRIPTION

Web interface

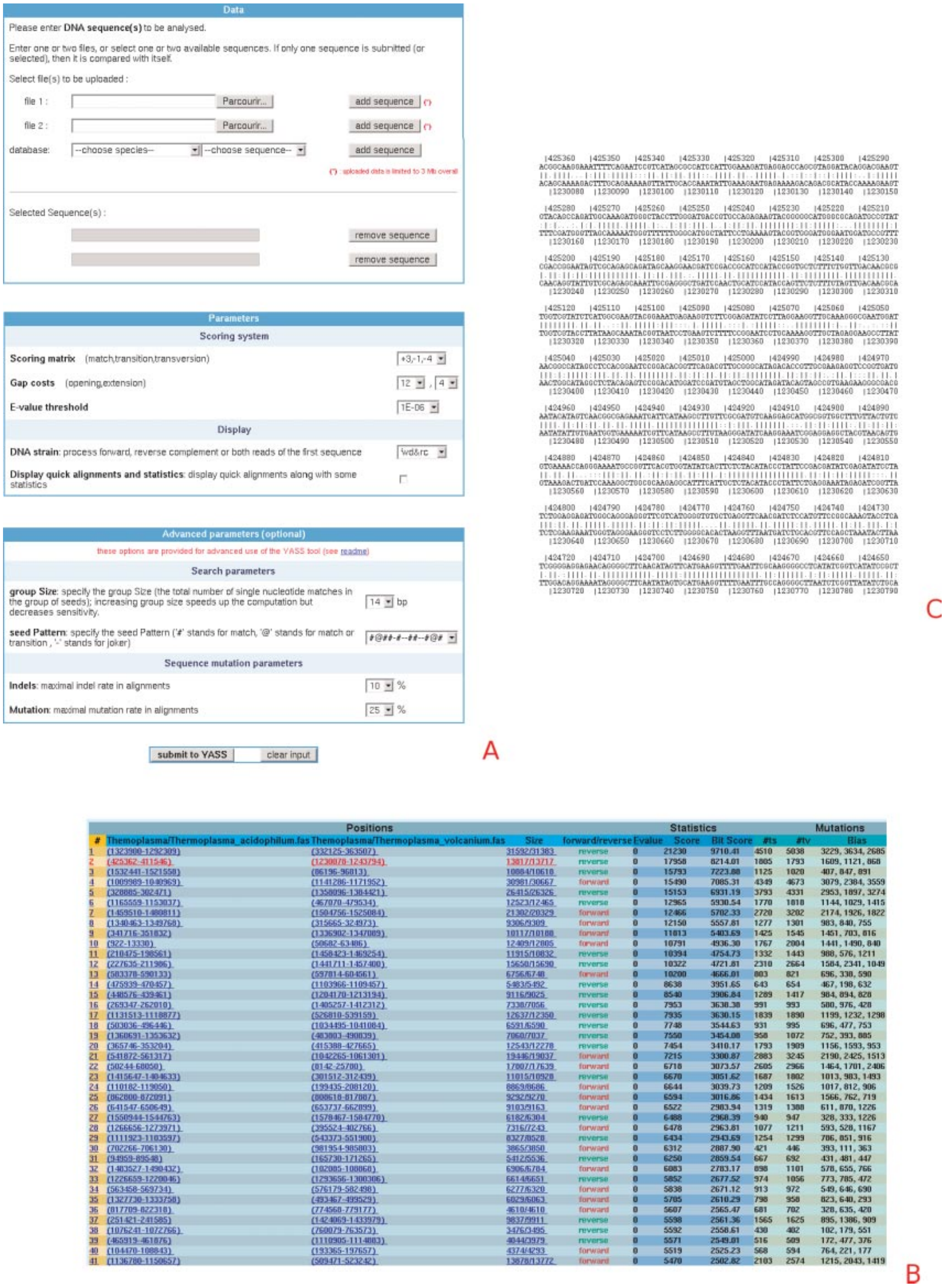
The main user input (Figure 1A) consists of one or two sequences in fasta format either chosen from a predefined database or uploaded to the web server.

Once sequences have been selected, the user can run the program right away with all other parameters set by default. Alternatively, the user can set other parameters such as the scoring matrix or gap penalties (preselected matrices are proposed), and specify the DNA strain to be processed (direct, complementary or both). The user can also choose to display complete alignments rather than only alignment positions.

More advanced parameters are available for expert users. For example, the right choice of the seed pattern can increase the search sensitivity considerably provided that some knowledge of target alignments is available (10–14). The web interface provides a preselection of seeds including three transition-constrained seeds, one providing a good performance compromise between coding and non-coding sequences, and the other two tuned respectively for non-coding and coding regions. The accompanying Hedera program (http://www.loria.fr/projects/YASS/hedera.html) is also provided for advanced users in order to design new seed patterns according to different probabilistic models of alignments (15).

Finally, the user can specify some statistical parameters of target alignments, such as the assumed substitution rate or indel rate. These parameters control the hit criterion, i.e. the rules for grouping together closely located seeds to detect similarities.

*To whom correspondence should be addressed. Tel: +33 3 83 59 30 11; Fax: +33 3 83 27 83 19; Email: Laurent.Noe@loria.fr



Data

Please enter **DNA sequence(s)** to be analysed.

Enter one or two files, or select one or two available sequences. If only one sequence is submitted (or selected), then it is compared with itself.

Select file(s) to be uploaded :

file 1 : (*)

file 2 : (*)

database: --choose species-- --choose sequence--

(*) : uploaded data is limited to 3 Mb overall

Selected Sequence(s) :

Parameters

Scoring system

Scoring matrix (match,transition,transversion): +3,-1,-4

Gap costs (opening,extension): 12 , 4

E-value threshold: 1E-06

Display

DNA strain: process forward, reverse complement or both reads of the first sequence: fwd&rc

Display quick alignments and statistics: display quick alignments along with some statistics ☐

Advanced parameters (optional)

these options are provided for advanced use of the YASS tool (see readme)

Search parameters

group Size: specify the group Size (the total number of single nucleotide matches in the group of seeds); increasing group size speeds up the computation but decreases sensitivity.: 14 bp

seed Pattern: specify the seed Pattern ('#' stands for match, '@' stands for match or transition , '-' stands for joker): #@##-#--##--#@#

Sequence mutation parameters

Indels: maximal indel rate in alignments: 10 %

Mutation: maximal mutation rate in alignments: 25 %

Positions			Statistics					Mutations		
#	Themoplasma/Thermoplasma_acidophilum.fas	Themoplasma/Thermoplasma_volcanium.fas	Size	forward/reverse	Evalue	Score	Bit Score	#ts	#tv	Bias
1	(1323900-1292309)	(332125-363507)	31592/31383	reverse	0	21230	9710.41	4510	5038	3229, 3634, 2685
2	(425362-411546)	(1230078-1243794)	13817/13717	reverse	0	17958	8214.01	1805	1793	1609, 1121, 868
3	(1532441-1521558)	(86196-96813)	10884/10618	reverse	0	15793	7223.88	1125	1020	407, 847, 891
4	(1009989-1040969)	(1141286-1171952)	30981/30667	forward	0	15490	7085.31	4349	4673	3079, 2384, 3559
5	(328885-302471)	(1358096-1384421)	26415/26326	reverse	0	15153	6931.19	3793	4331	2953, 1897, 3274
6	(1165559-1153037)	(467070-479534)	12523/12465	reverse	0	12965	5930.54	1770	1818	1144, 1029, 1415
7	(1459510-1480811)	(1504756-1525084)	21302/20329	forward	0	12466	5702.33	2720	3202	2174, 1926, 1822
8	(1340463-1349768)	(315665-324973)	9306/9309	forward	0	12150	5557.81	1277	1301	983, 840, 755
9	(341716-351832)	(1336902-1347089)	10117/10188	forward	0	11813	5403.69	1425	1545	1451, 703, 816
10	(922-13330)	(50682-63486)	12409/12805	forward	0	10791	4936.30	1767	2004	1441, 1490, 840
11	(210475-198561)	(1458423-1469254)	11915/10832	reverse	0	10394	4754.73	1332	1443	988, 576, 1211
12	(227635-211986)	(1441711-1457400)	15650/15690	reverse	0	10322	4721.81	2310	2664	1584, 2341, 1049
13	(583378-590133)	(597814-604561)	6756/6748	forward	0	10200	4666.01	803	821	696, 338, 590
14	(475939-470457)	(1103966-1109457)	5483/5492	reverse	0	8638	3951.65	643	654	467, 198, 632
15	(448576-439461)	(1204170-1213194)	9116/9025	reverse	0	8540	3906.84	1289	1417	984, 894, 828
16	(269347-262010)	(1405257-1412312)	7338/7056	reverse	0	7953	3638.38	991	993	580, 976, 428
17	(1131513-1118877)	(526810-539159)	12637/12350	reverse	0	7935	3630.15	1839	1890	1199, 1232, 1298
18	(503036-496446)	(1034495-1041084)	6591/6590	reverse	0	7748	3544.63	931	995	696, 477, 753
19	(1360691-1353632)	(483803-490839)	7060/7037	reverse	0	7550	3454.08	958	1072	752, 393, 885
20	(365746-353204)	(415388-427665)	12543/12278	reverse	0	7454	3410.17	1793	1909	1156, 1593, 953
21	(541872-561317)	(1042265-1061301)	19446/19037	forward	0	7215	3300.87	2883	3245	2190, 2425, 1513
22	(50244-68050)	(8142-25780)	17807/17639	forward	0	6718	3073.57	2605	2966	1464, 1701, 2406
23	(1415647-1404633)	(301512-312439)	11015/10928	reverse	0	6670	3051.62	1687	1802	1013, 983, 1493
24	(110182-119050)	(199435-208120)	8869/8686	forward	0	6644	3039.73	1209	1526	1017, 812, 906
25	(862800-872091)	(808618-817887)	9292/9270	forward	0	6594	3016.86	1434	1613	1566, 762, 719
26	(641547-650649)	(653737-662899)	9103/9163	forward	0	6522	2983.94	1319	1388	611, 870, 1226
27	(1550944-1544763)	(1578467-1584770)	6182/6304	reverse	0	6488	2968.39	940	947	328, 333, 1226
28	(1266656-1273971)	(395524-402766)	7316/7243	forward	0	6478	2963.81	1077	1211	593, 528, 1167
29	(1111923-1103597)	(543373-551900)	8327/8528	reverse	0	6434	2943.69	1254	1299	706, 851, 916
30	(702266-706130)	(981954-985803)	3865/3850	forward	0	6312	2887.90	421	446	393, 111, 363
31	(94959-89548)	(165730-171265)	5412/5536	reverse	0	6250	2859.54	667	692	431, 481, 447
32	(1483527-1490432)	(102085-108068)	6906/6784	forward	0	6083	2783.17	898	1101	578, 655, 766
33	(1226659-1220046)	(1293656-1300306)	6614/6651	reverse	0	5852	2677.52	974	1056	773, 785, 472
34	(563458-569734)	(576179-582498)	6277/6320	forward	0	5838	2671.12	913	972	549, 646, 690
35	(1327730-1333758)	(493467-499529)	6029/6063	forward	0	5705	2610.29	798	958	823, 640, 293
36	(817709-822318)	(774568-779177)	4610/4610	forward	0	5607	2565.47	681	702	328, 635, 420
37	(251421-241585)	(1424069-1433979)	9837/9911	reverse	0	5598	2561.36	1565	1625	895, 1386, 909
38	(1076241-1072766)	(760079-763573)	3476/3495	reverse	0	5592	2558.61	430	402	102, 179, 551
39	(465919-461876)	(1110905-1114883)	4044/3979	reverse	0	5571	2549.01	516	509	172, 477, 376
40	(104470-108843)	(193365-197657)	4374/4293	forward	0	5519	2525.23	568	594	764, 221, 177
41	(1136780-1150657)	(509471-523242)	13878/13772	forward	0	5470	2502.82	2103	2574	1215, 2043, 1419

```
 |425360   |425350   |425340   |425330   |425320   |425310   |425300   |425290
ACGGCAAGGAAATTTTCAGAATCCGTCATAGCGCCATCCATTGGAAAGATGAGGAGCCAGCGTAGGATACAGGACGAAGT
[illegible]
ACAGCAAAAGACTTTGCAGAAAAAGTTATTGCACCAAATATTGAAAGAATGAGAAAAGACAGACGCATACCAAAAGAAGT
  |1230080  |1230090  |1230100  |1230110  |1230120  |1230130  |1230140  |1230150

 |425280   |425270   |425260   |425250   |425240   |425230   |425220   |425210
GTACAGCCAGATGGCAAAGATGGGCTACCTTGGGATGACCGTGCCAGAGAAGTACGGGGGCATGGGCGCAGATGCCGTAT
[illegible]
TTTCGATGGGTTAGCAAAAATGGGTTTTTTGGGCATGGCTATTCCTGAAAAGTACGGTGGGATGGGAATGGATGCCGTTT
  |1230160  |1230170  |1230180  |1230190  |1230200  |1230210  |1230220  |1230230

 |425200   |425190   |425180   |425170   |425160   |425150   |425140   |425130
CGACGGGAATAGTCGCAGAGCAGATAGCAAGGAACGATCCGACCGCATCCATACCGGTGCTCTTTCTGGTTGACAACGGG
[illegible]
CAACAGGTATTGTCGCAGAGCAAATTGCGAGGGCTGATCCAACTGCATCCATACCAGTTCTCTTTCTAGTTGACAACGCA
  |1230240  |1230250  |1230260  |1230270  |1230280  |1230290  |1230300  |1230310

 |425120   |425110   |425100   |425090   |425080   |425070   |425060   |425050
TGGTCGTATCTCATGGCGAAGTACGGAAATGAGAAGGTCTTCGGAGATATCCTTAGGAAGGTTGCAAAGGGCGAATGGAT
[illegible]
TGGTCGTACCTTATAAGCAAATACGGTAATCCTGAAGTCTTTTCCGGAATCCTGCAAAAGGTTGCTAGAGGAAGCCTTAT
  |1230320  |1230330  |1230340  |1230350  |1230360  |1230370  |1230380  |1230390

 |425040   |425030   |425020   |425010   |425000   |424990   |424980   |424970
AACGGCCATAGCCTCCACGGAATCCGGACACGGTTCAGACGTTGCGGGCATAGACACCGTTGCGAAGAGGTCCGGTGATG
[illegible]
AACTGGCATAGGCTCTACAGAGTCCGGACATGGATCCGATGTAGCTGGCATAGATACAGTAGCCGTGAAGAAGGGCGACG
  |1230400  |1230410  |1230420  |1230430  |1230440  |1230450  |1230460  |1230470

 |424960   |424950   |424940   |424930   |424920   |424910   |424900   |424890
AATACATAGTCAACGGCGAGAAATCATTCATAAGGCTTGTTCGCCGATGTCAAGGAGCATGGGGTGGCTTTGTTACTGTC
[illegible]
AATATATTGTGAATGGTGAAAAATCGTTCATAAGGCTTGTAAGGGATATCAAGGAAATCGGAGGAGGCTACGTAACAGTG
  |1230480  |1230490  |1230500  |1230510  |1230520  |1230530  |1230540  |1230550

 |424880   |424870   |424860   |424850   |424840   |424830   |424820   |424810
GTGAAAACCAGGGAAAATGCCGGTTCACGTGGTATATCACTTCTCTACATACCCTATTCCGACGATATCGAGATATCCTA
[illegible]
GTAAAGACTGATCCAAAGGCTGGCGCAAGAGGCATTTCATTGCTCTACATACCCTATTCTGAGGAAATAGAGATCGGTTA
  |1230560  |1230570  |1230580  |1230590  |1230600  |1230610  |1230620  |1230630

 |424800   |424790   |424780   |424770   |424760   |424750   |424740   |424730
TCTGGAGGAGATGGGCAGGGAGGGTTCGTCATGGGGTGTGCTGAGGTTCAACGATCTCCATGTTCCGGCAAAGTACCTCA
[illegible]
TCTCGAAGAAATGGGTAGGGAAGGGTCCTCTTGGGGCACACTAAGGTTTAATGATCTGCACGTTCCAGCTAAATACTTAA
  |1230640  |1230650  |1230660  |1230670  |1230680  |1230690  |1230700  |1230710

 |424720   |424710   |424700   |424690   |424680   |424670   |424660   |424650
TCGGGGAGGAGAACAGGGGCTTCAACATAGTTCATGAAGGTTTTGAATTCGCAAGGGGCCTCATATCGGTCATATCCGCT
[illegible]
TTGGACAGGAAAATAGGGGCTTCAATATAGTGCATGAAGGTTTTGAATTTGCCAGGGGCTTAATCTCGGTTATATCTGCA
  |1230720  |1230730  |1230740  |1230750  |1230760  |1230770  |1230780  |1230790
```

Figure 1. The input window (**A**) allows users to control most YASS parameters, from most basic to more advanced. Results are output in tabular format (**B**), with the possibility of displaying each sequence alignment (**C**).

Once the results are obtained, it is possible to generate a clickable dot-plot (Figure 2) where each alignment is linked to a URL with its text representation (Figure 1C). A tabular output (Figure 1B) is also available: alignments are sorted according to their *E*-value and linked to their text representation. Finally, the YASS output can also be downloaded in text format for further analysis.

Technical issues. The YASS server available at http://yass.loria.fr/interface.php currently runs Apache 2.0.47

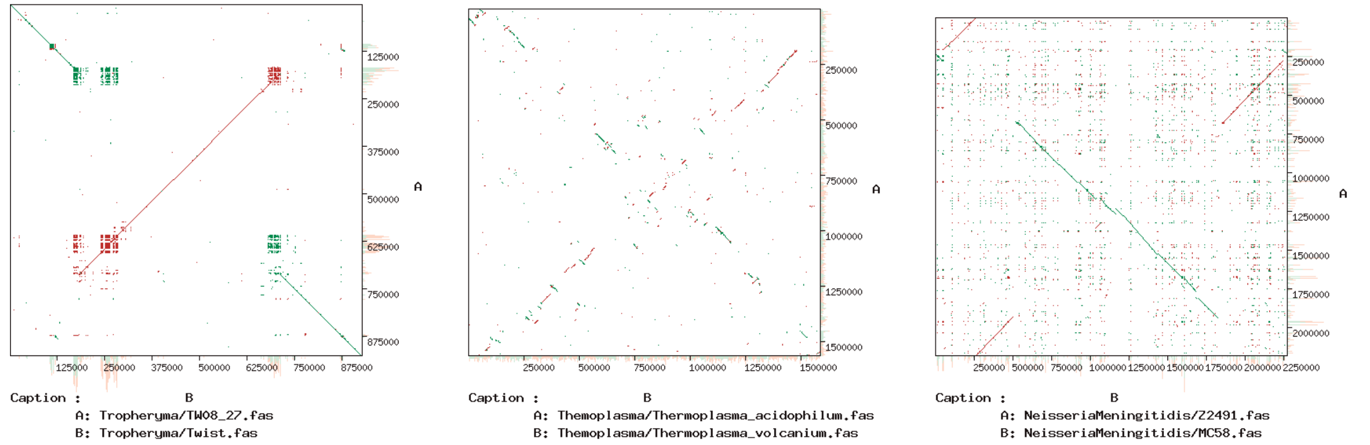


Figure 2. Three YASS dot-plots are shown, each obtained from pairs of closely related bacterial sequences. Green segments represent alignments of forward reads and red segments correspond to alignments between the reverse complement of one sequence and the forward read of the other.

(PHP and Perl-CGI modules) on a Linux Mandrake 9.2. Dot-plots are obtained with the GD graphical library interfaced to PHP. The YASS program has been developed in C and is distributed under the Gnu General Public License.

Owing to limitations of computational resources, some restrictions have been made on the web interface. For example, uploaded files are currently limited to 3 Mb, scoring systems can be chosen only among preselected ones and for each parameter a fixed range of possible values has been settled.

Standalone version

The standalone version is recommended for frequent users or those who need specific parameters to be set outside preselected values. It provides access to two other output formats, including a BLAST-like tabular format that can be used by existing postprocessing parsers. Note that YASS does not need one of the sequences to be preprocessed (*formatdb* command of BLAST), rather, it treats both sequences on the fly.

METHODS

Here we briefly outline the underlying principles of the YASS algorithm, including some novel features. For a more detailed presentation the reader is referred to (16) (http://www.biomedcentral.com/1471-2105/5/149/).

Seed model

Seeds are specified using a seed pattern built over a three-letter alphabet #, @ and –, where # stands for a nucleotide match, – for a don't care symbol and @ for a match or a transition (mutation A↔G or C↔T). The weight of a pattern is defined as the number of # plus half the number of @. The weight is the main characteristic of seed selectivity.

The advantage of transition-constrained seeds stems from the biological observation that transition mutations are relatively more frequent than transversions, in both coding and non-coding regions. Typically, biologically relevant alignments contain about the same number of transitions and transversions, whereas transitions are half as frequent in independently and identically distributed random sequences. Transition-constrained seeds increase the possible number of transitions in a hit relative to spaced seeds without the transition constraint, and this is done without loss of sensitivity or efficiency.

The sensitivity of a given seed has been estimated using the algorithm of (15), which is a generalization of the one proposed in (11). Two main alignment models have been considered: a Bernoulli model (13) assumed to simulate alignments of non-coding DNA and a hidden Markov model (10) assumed to simulate alignments of coding DNA. By default, YASS currently uses the seed #@# −− ## −− # − ##@# of weight 9, which provides a good compromise in detecting similarities in both coding and non-coding sequences. The standalone version of YASS allows users to specify their own seeds. Several preselected seeds are provided by the YASS web interface.

Hit criterion

YASS is based on a multi-seed hit criterion that defines a hit as a group of closely located and possibly overlapping seeds. Two seeds belong to the same group if they occur within a bounded distance or, on the other hand, are located at close dot-plot diagonals. Distance threshold parameters are computed according to probabilistic sequence models taking into account substitution and indel rates, similarly to models used in (17). Note that seeds of a group are allowed to overlap. An additional group size parameter sets a lower bound on the total number of individual matches and transitions of the group. Using the group size results in a flexible criterion that combines a guaranteed selectivity with a good sensitivity on both short and long similarities. More details on the hit criterion can be found in (16).

Comparative tests

To validate the better performance of transition-constrained seeds compared with ordinary spaced seeds, several comparative experiments have been presented in (16). Transition-constrained seeds have been shown to be more sensitive with respect to some Bernoulli and hidden Markov models of alignments of coding and non-coding DNA [Tables 1 and 2 in

(16)]. Moreover, transition-constrained seeds have been shown to be more sensitive in detecting alignments of real genomic sequences [Table 3 in (16)].

YASS has been compared with bl2seq (NCBI BLAST 2.2.6) according to several criteria: running time, number of significant alignments found (with E-value $\leqslant 10^{-6}$) and number of significant alignments found exclusively by one program and their total length [Table 4 in (16)]. The results show that YASS detects more significant alignments than bl2seq, within a smaller time for large DNA sequences.

CONCLUSIONS

In this paper, we have described YASS—a new DNA local alignment tool. The proposed web interface features several output formats suitable for a *coup d'oeil* analysis as well as for a deeper analysis of alignments. An upcoming release of YASS will include multi-seed indexing strategies and an optimized processor-cache algorithm.

ACKNOWLEDGEMENTS

The authors wish to thank Marie-Pierre Etienne and Pierre Valois for their help during early stages of this work, Mikhail Roytberg for his contribution to the seed design problem and Christophe Valmir for improving the YASS code. The authors have been supported by the ACI IMPBio of the French Ministry of Research. They thank the reviewers for their helpful remarks. Funding to pay the Open Access publication charges for this article was provided by CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
2. Lipman,D. and Pearson,W. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
3. Califano,A. and Rigoutsos,I. (1993) Flash: a fast look-up algorithm for string homology. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1**, 56–64.
4. Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
5. Altschul,S., Madden,T., Schäffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
6. Schwartz,S., Kent,J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
7. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
8. Burkhardt,S. and Kärkkäinen,J. (2001) Better filtering with gapped q-grams. In *Proceedings of the 12th Symposium on Combinatorial Pattern Matching (CPM'01)*, 1–4 July, Jerusalem, Israel, *LNCS* 2089, Springer, pp. 73–85.
9. Li,M., Ma,B., Kisman,D. and Tromp,J. (2004) PatternHunter II: highly sensitive and fast homology search. *J. Bioinform. Comput. Biol.*, **2**, 417–439.
10. Brejova,B., Brown,D. and Vinar,T. (2003) Optimal spaced seeds for hidden Markov models, with application to homologous coding regions. In Baeza-Yates,R., Chavez,E. and Crochemore,M. (eds), *Proceedings of the 14th Symposium on Combinatorial Pattern Matching*, 25–27 June, Morelia, Mexico, *LNCS* 2676, Springer, pp. 42–54.
11. Buhler,J., Keich,U. and Sun,Y. (2003) Designing seeds for similarity search in genomic DNA. In *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB'03)*, 10–13 April, Berlin, Germany, ACM Press, pp. 67–75.
12. Choi,K.P., Zeng,F. and Zhang,L. (2004) Good spaced seeds for homology search. *Bioinformatics*, **20**, 1053–1059.
13. Keich,U., Li,M., Ma,B. and Tromp,J. (2004) On spaced seeds for similarity search. *Discrete Appl. Math.*, **138**, 253–263.
14. Kucherov,G., Noé,L. and Ponty,Y. (2004) Estimating seed sensitivity on homogeneous alignments. In *Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE2004)*, May 19–21, Taichung, Taiwan, IEEE Computer Society Press, pp. 387–394.
15. Kucherov,G., Noé,L. and Roytberg,M. (2004) A unifying framework for seed sensitivity and its application to subset seeds. Rapport de recherche INRIA RR-5374. http://www.inria.fr/rrrt/rr-5374.html.
16. Noé,L. and Kucherov,G. (2004) Improved hit criteria for DNA local alignment. *BMC Bioinformatics*, **5**, 149.
17. Benson,G. (1999) Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.